

project_proposal

November 12, 2025

1 Project Proposal Title:

NASS Crop Yield Predictor and Analysis for Texas farms

1.1 Teammates

- Stephen Cox, ssj63
- Shazz Momin, wzu2

1.2 Project Abstract

The goal of this project is to use past agricultural census to build a model that can predict crop yield for Texas farms and explore the meaning of what our model predicts

1.3 Problem Statement

- Farmers in Texas are a very important part of our community. It is important to know what factors can effect the success of our farmers and their product since many people relay on Texas grown crops.
- What factors most deeply effect our farms success and can we predict if our farms will be succesful for the year.
- Benchmarks we will use are Mean Absolute Error, Root Mean Squared Error and R^2 . Each of these scores we help us interpret different parts of the models accuracy/precision/recall since this is a regressive model
- The Data comes from US Department of Agriculture - National Agricultural Statistics Service census data which includes many different info points like: yield, land size and fertilizer...
- Practical Interpretability, we hope to use this model to try and interpret correlations between farming factors and crop yield
- What we hope to achieve
 - Build a working NN regressor that predicts county-level crop yield for Texas using 2012/2017/2022 data.
 - Beat simple baselines (mean and linear reg.) by at least a measurable margin (lower MAE/RMSE).
 - Produce interpretable model explainers (SHAP or partial dependence) showing the most influential

```
[ ]: import pandas as pd

file_path = "data/Texas_AgCensus_2012_2017_2022.csv"
chunksize = 200_000
total_floats = 0

for chunk in pd.read_csv(file_path, chunksize=chunksize, low_memory=False):
    num_cols = chunk.select_dtypes(include=['float64', 'int64'])
    total_floats += num_cols.count().sum()

print(f"Total float/int entries: {total_floats:,}")
print("Total float/int entries > 10,000,000")
```

Total float/int entries: 12,723,963
Total float/int entries > 10,000,000

```
[20]: df = pd.read_csv("data/Texas_AgCensus_2012_2017_2022.csv")
df
```

C:\Users\scox1\AppData\Local\Temp\ipykernel_11356\167767244.py:1: DtypeWarning:
Columns (18,21,29,37,38) have mixed types. Specify dtype option on import or set
low_memory=False.

```
df = pd.read_csv("data/Texas_AgCensus_2012_2017_2022.csv")
```

```
[20]:
```

	SOURCE_DESC	SECTOR_DESC	GROUP_DESC \
0	CENSUS	ANIMALS & PRODUCTS	LIVESTOCK
1	CENSUS	CROPS	FRUIT & TREE NUTS
2	CENSUS	ECONOMICS	FARMS & LAND & ASSETS
3	CENSUS	ANIMALS & PRODUCTS	LIVESTOCK
4	CENSUS	ANIMALS & PRODUCTS	POULTRY
...
1240152	CENSUS	DEMOGRAPHICS	OPERATORS
1240153	CENSUS	DEMOGRAPHICS	OPERATORS
1240154	CENSUS	DEMOGRAPHICS	OPERATORS
1240155	CENSUS	ANIMALS & PRODUCTS	LIVESTOCK
1240156	CENSUS	ANIMALS & PRODUCTS	LIVESTOCK

	COMMODITY_DESC	CLASS_DESC \
0	CATTLE	(EXCL COWS)
1	GRAPES	ALL CLASSES
2	AG LAND	CROPLAND, HARVESTED
3	CATTLE	COWS, BEEF
4	EMUS	ALL CLASSES
...
1240152	OPERATORS, PRINCIPAL	AGE 45 TO 54
1240153	OPERATORS, PRINCIPAL	AGE 55 TO 64
1240154	OPERATORS, PRINCIPAL	AGE GE 65
1240155	CATTLE	ALL CLASSES

1240156 CATTLE ALL CLASSES

	PRODN_PRACTICE_DESC	UTIL_PRACTICE_DESC \
0	ALL PRODUCTION PRACTICES	ALL UTILIZATION PRACTICES
1	ALL PRODUCTION PRACTICES	ALL UTILIZATION PRACTICES
2	ALL PRODUCTION PRACTICES	ALL UTILIZATION PRACTICES
3	ALL PRODUCTION PRACTICES	ALL UTILIZATION PRACTICES
4	ALL PRODUCTION PRACTICES	ALL UTILIZATION PRACTICES
...
1240152	PRIMARY OCCUPATION, (EXCL FARMING)	ALL UTILIZATION PRACTICES
1240153	PRIMARY OCCUPATION, (EXCL FARMING)	ALL UTILIZATION PRACTICES
1240154	PRIMARY OCCUPATION, (EXCL FARMING)	ALL UTILIZATION PRACTICES
1240155	ON FEED	ALL UTILIZATION PRACTICES
1240156	ON FEED	ALL UTILIZATION PRACTICES

	STATISTICCAT_DESC	UNIT_DESC \
0	INVENTORY	OPERATIONS
1	AREA BEARING	ACRES
2	AREA	ACRES
3	INVENTORY	OPERATIONS
4	SALES	OPERATIONS
...
1240152	AGE, AVG	YEARS
1240153	AGE, AVG	YEARS
1240154	AGE, AVG	YEARS
1240155	INVENTORY	OPERATIONS
1240156	INVENTORY	HEAD

	SHORT_DESC ... YEAR \
0	CATTLE, (EXCL COWS) - OPERATIONS WITH INVENTORY ... 2022
1	GRAPES - ACRES BEARING ... 2022
2	AG LAND, CROPLAND, HARVESTED - ACRES ... 2022
3	CATTLE, COWS, BEEF - OPERATIONS WITH INVENTORY ... 2022
4	EMUS - OPERATIONS WITH SALES ... 2022
...	...
1240152	OPERATORS, PRINCIPAL, AGE 45 TO 54, PRIMARY OC... ... 2012
1240153	OPERATORS, PRINCIPAL, AGE 55 TO 64, PRIMARY OC... ... 2012
1240154	OPERATORS, PRINCIPAL, AGE GE 65, PRIMARY OCCUP... ... 2012
1240155	CATTLE, ON FEED - OPERATIONS WITH INVENTORY ... 2012
1240156	CATTLE, ON FEED - INVENTORY ... 2012

	FREQ_DESC	BEGIN_CODE	END_CODE	REFERENCE_PERIOD_DESC \
0	POINT IN TIME	12	12	END OF DEC
1	ANNUAL	0	0	YEAR
2	ANNUAL	0	0	YEAR
3	POINT IN TIME	12	12	END OF DEC
4	ANNUAL	0	0	YEAR

```

...
1240152 POINT IN TIME      12      12      END OF DEC
1240153 POINT IN TIME      12      12      END OF DEC
1240154 POINT IN TIME      12      12      END OF DEC
1240155 POINT IN TIME      12      12      END OF DEC
1240156 POINT IN TIME      12      12      END OF DEC

```

```

      WEEK_ENDING      LOAD_TIME  VALUE  CV_%  CENSUS_YEAR
0      NaN  2024-02-13 12:00:00     35    (L)         2022
1      NaN  2024-02-13 12:00:00    243    (L)         2022
2      NaN  2024-02-13 12:00:00    348    (L)         2022
3      NaN  2024-02-13 12:00:00     26   36.9         2022
4      NaN  2024-02-13 12:00:00      1    (L)         2022
...
1240152      NaN  2012-12-31 00:00:00   50.1    (L)         2012
1240153      NaN  2012-12-31 00:00:00   59.3    (L)         2012
1240154      NaN  2012-12-31 00:00:00   73.3    0.1         2012
1240155      NaN  2012-12-31 00:00:00    90   14.1         2012
1240156      NaN  2012-12-31 00:00:00  1,335   13.2         2012

```

[1240157 rows x 40 columns]

```
[19]: df.shape
```

```
[19]: (1240157, 40)
```

1.4 Dataset

- The Data set is a derivative we made from NASS 2022, 2017 and 2012 census filtered for Texas. The total floats is greater than 12,723,000 floats.
- The NASS data set includes many different farming insights like the amount of fertilizer used or land size as well as crop yields for specific farms. The number of extra information will help make a more precise model with the colinearity involved.
- The size of the data set is (1240157, 40) each instance is a different measure for a farm. The columns include the description of what's measure in the column short_desc and many other columns help catagorize the data

1.5 Methodology

- The Baseline model will be Random Forest Regressor and Linear Regression and the improvement will be a Neural Network Regressor. We will be using libraries like sklearn, pandas, NumPy, seaborn, matplotlib, tensorflow and tensorboard

1.6 Teaming Strategy

We will split the project into seperate main sections

Stephen: - dataset preparation - improvement model building and testing - improvement model visualization

Shazz: - baseline model building and testing - baseline model visualization

Together: - analysis of baseline vs improvement - extended analysis on improvement model and its implications - presentation of models and analysis

We will try to meet at least once a week and use discord as a communication platform for progress

1.6.1 Role Assignments and Commitment Matrix

- For the project progress will have the models built and tested, after project progress we will work on the analysis and making of the slides
- To Collaborate we will be using GitHub and discord to share code and communicate

1.7 Mitigation Plan

- IF the data is too inconsistent and NN fails to converge we will use a simpler improvement model or reduce the scale from State level to County
- In case of MIA we will focus on the essentials of the project, primarily the baseline models, visualization and presentation.
- If Baseline data is GIGO we will reassess the data preprocessing procedure, and try to simplify the data